The Annals of Statistics
1999, Vol. 27, No. 1, 129–142

CONVERGENCE AND ACCURACY OF GIBBS SAMPLING FOR CONDITIONAL DISTRIBUTIONS IN GENERALIZED LINEAR MODELS[1]

BY JOHN E. KOLASSA

University of Rochester Medical Center

This paper presents convergence conditions for a Markov chain constructed using Gibbs sampling, when the equilibrium distribution is the conditional sampling distribution of sufficient statistics from a generalized linear model. For cases when this unidimensional sampling is done approximately rather than exactly, the difference between the target equilibrium distribution and the resulting equilibrium distribution is expressed in terms of the difference between the true and approximating univariate conditional distributions. These methods are applied to an algorithm facilitating approximate conditional inference in canonical exponential families.

1. Introduction. This paper presents conditions for convergence of an algorithm that simulates observations from distributions approximating null conditional distributions in generalized linear models in order to construct multivariate conditional significance tests and confidence regions. Of interest are models in which independent responses Y_j are observed from a density or mass function of the form

$$\exp\big(\eta_j y_j - L(\eta_j) - h(y_j)\big) \quad \text{for } \eta_j = \mathbf{z}_j \beta \text{ and } y_j \in \mathfrak{Y}_j \tag{1}$$

with row vectors of covariates $\mathbf{z}_j \in \mathbb{R}^d$ forming rows of a full-rank matrix $\mathbf{Z}$. Typically β may take any value in $\mathbb{R}^d$. Applications in this paper will generally have $\mathfrak{Y}_j \subset \mathbb{Z}$, or $\mathfrak{Y}_j$ a connected subset of $\mathbb{R}$. Sufficient statistics are of the form

$$\mathbf{T} = \mathbf{Z}^{\mathsf{T}}\mathbf{Y}, \quad \text{with } \mathbf{Y} \in \prod_j \mathfrak{Y}_j. \tag{2}$$

Of interest is inference on components $\{1, \dots, a\}$ of β without specification of the remaining components. Hypothesis testing in this context is performed by enumerating the sample space and associated probabilities of $(T_1, \dots, T_a)$, conditional on $(T_{a+1}, \dots, T_d)$. These probabilities are

$$\left[\sum_{\mathbf{y} \in \mathscr{Y}(\mathbf{t})} \exp\Big(-\sum_j h(y_j)\Big)\right]\exp(\ell(\beta;\mathbf{t})) \quad \text{for } \ell(\beta;\mathbf{t}) = \beta^{\mathsf{T}}\mathbf{t} - \sum_j L\big(\mathbf{z}_j^{\mathsf{T}}\beta\big). \tag{3}$$

for $\mathscr{Y}(\mathbf{t}) = \{\mathbf{y} \mid \mathbf{z}_{a+1}^{\mathsf{T}}\mathbf{y} = t_{a+1}, \dots, \mathbf{z}_d^{\mathsf{T}}\mathbf{y} = t_d\}$, where $\mathbf{z}_i$ is row i of $\mathbf{Z}$. Were exploration of a posterior distribution on β desired, a Metropolis–Hastings

Received April 1997; revised August 1998.

[1] Supported by NIH Grants CA 63050 and CA 35464.

AMS 1991 *subject classifications*. Primary 60J20; secondary 62E20.

Key words and phrases. Markov chain Monte Carlo, saddlepoint approximations.

chain might be constructed by sampling from an approximation to (3), and correcting this by evaluating (3) directly. In such cases, the quantity in square brackets in (3) need not be evaluated.

Sampling values of $\mathbf{t}$, rather than values of β, however, requires calculating the summation in (3). Algorithms exist for such calculations for a logistic regression [Hirji, Mehta and Patel (1987)]. These calculations may generally only be done for sufficiently small data sets; for larger data sets many authors have recourse to Monte Carlo methods [Mehta, Patel and Senchaudhuri (1993), Forster, McDonald and Smith (1996), Diaconis and Sturmfels (1998)]. None of these authors has found evaluation of the bracketed quantity of (3) practical in conjunction with Monte Carlo methods.

Instead, probabilities in (3) will be calculated by using the Gibbs sampler to construct a Markov chain whose null distribution approximates the conditional distribution of interest. This approximating chain is constructed by sampling from an asymptotic approximation instead of the true conditional cumulative distribution functions. This paper provides a theoretical justification for this algorithm in three ways: first, reducibility of both chains is assessed. Next, a simple proof is given demonstrating that when the constructed Markov chains are irreducible, they have equilibrium distributions. Finally, the resulting equilibrium distribution is shown to approximate the desired conditional distribution to order $O(1/n)$, where n indexes the number of independent replicates represented in the data set.

In cases in which the operator representing transformations has norm strictly less than one, Schervish and Carlin (1992) approach similar problems by representing the distribution at each step in the chain as a member of a Hilbert space of distributions. Roberts and Polson (1994) address convergence questions for operators with a different norm bounded away from unity. These norm conditions do not hold for chains examined here. This paper makes use of Markov chain convergence results in the more general cases; these methods are described by Nummelin (1984) and reviewed by Tierney (1994). Furthermore, these methods address convergence in cases where the sampling performed is according to the Gibbs scheme, except that this sampling is not from the exact conditionals, but from approximations to these conditions. Roberts and Smith (1994) discuss conditions of aperiodicity and irreducibility necessary for convergence.

The first section below presents background on methods used. The second formally defines the Markov chain and presents results on its convergence properties.

2. Markov chain and distribution function approximation background.

This section reviews background on Markov chain terminology, Gibbs sampling and conditional distribution function approximations.

2.1. *Markov chain terminology.*

To make connections between this work and other Markov chain literature clearer, some common definitions concerning Markov chains are introduced. We first consider whether the state space may be divided into two spaces, between which the chain never travels. If this is the case, the chain is called reducible, and there are an infinite number of equilibrium distributions for the chain, depending on how much mass is initially allocated to each subspace. Each simulated Markov chain will remain in one division of the state space. If multiple chains are simulated, with starting points selected independently from some distribution, whether the proportion of time the chain visits a subset of the sample space (after a number of initial iterations necessary to approximate equilibrium) is a reasonable estimate of the probability of that set depends on whether the correct proportion of chains was started in each noncommunicating subset. Ensuring that this happens is generally impossible. Hence we restrict attention to irreducible chains.

We require a definition concerning whether the measure induced by certain transitions in the chain can be bounded below by a measure that does not depend on the initial state; many convergence results depend on this property. This section concludes with the main Markov chain convergence result needed in this paper.

DEFINITION 2.1. A set $\mathscr{T} \subset \mathfrak{X}$ is small if and only if there exists a constant $\alpha > 0$ and a probability measure ν on $\mathfrak{X}$ such that $P(\mathbf{t}, \cdot\,) \geq \alpha\nu(\cdot)\ \forall\ \mathbf{t} \in \mathscr{T}$.

DEFINITION 2.2. A Markov chain is geometrically ergodic, if there exists a probability measure π, a constant $\omega > 1$ and a function $\psi(\mathbf{v})$ such that $\|P^{(m)}(\mathbf{v}, \cdot\,) - \pi(\cdot)\|_{\mathrm{TV}} < \psi(\mathbf{v})\omega^{-m}$. Here $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm on the space of finite measures on $\mathfrak{X}$.

Nummelin [(1984), Theorem 6.14] gives other logically equivalent definitions of geometric ergodicity.

Convergence properties of the Markov chain are governed by the size of iterates of the transition density, after the contributions of the small sets are removed. Rosenthal (1995) extends results of Nummelin (1984) on geometric convergence of Markov chains to provide a bound on the convergence rates for the associated chains. For the present purposes the following summary of this result will suffice.

LEMMA 2.3. *Suppose that* $\{\mathbf{T}_n^{(m)}\}$ *is a family of Markov chains indexed by* n, *on a common state space* $\mathfrak{X}^*$, *with small sets* $\mathscr{T}_n$, *associated measures* ν_n *and parameters* α_n. *Suppose that* α_n *and* $\theta_n = \int_{\mathfrak{X}^*} \nu_n(d\mathbf{t}) s_n(\mathbf{t})$ *are bounded away from zero. Suppose that there exists positive measurable functions* g_n *on* $\mathfrak{X}^*$ *and constants* ω_{0n}, $\gamma_n > 0$ *and* Γ_n *such that* $\omega_{0n} > 1$ *is bounded away from* 1 *and*

$$
\begin{aligned}
&\sup_{\mathbf{t} \notin \mathscr{T}} \mathrm{E}\big[\, \omega_{0n} g_n(\mathbf{T}^{(m+1)}) - g_n(\mathbf{T}^{(m)}) \mid \mathbf{T}^{(m)} = \mathbf{t}\big] \leq -\gamma_n \quad \text{for } \gamma_n, \\
&\sup_{\mathbf{t} \in \mathscr{T}_n} \mathrm{E}\big[\, g_n(\mathbf{T}^{(m+1)})\big(1 - s_n(\mathbf{T}^{(m+1)})\big) \mid \mathbf{T}^{(m)} = \mathbf{t}\big] < \Gamma_n \gamma_n < \infty.
\end{aligned} \tag{4}
$$

Suppose further that the associated parameters γ_n, Γ_n and ω_{0n} are bounded. Then for each n an equilibrium measure π_n and a function ψ_n exist, and a constant $\omega > 0$ exists, satisfying $\|P_n^{(m)}(\mathbf{v}, \cdot\,) - \pi_n(\cdot)\|_{\mathrm{TV}} < \psi(\mathbf{v})\omega^{-m}$ for all n. Furthermore, the integrals $\int_{\mathfrak{X}^*}\psi_n(\mathbf{v})\pi_n(d\mathbf{v})$ are uniformly bounded.

2.2. *Gibbs sampling.* The Gibbs sampler is a popular Markov chain method useful for yielding a sample from a posterior or likelihood density. It was first introduced by Geman and Geman (1984) in the context of image reconstruction. The data augmentation algorithm of Tanner and Wong (1987), introduced as a device for the calculation of posterior distributions, is a two-component version of the Gibbs sampler. See Tanner (1996) for background details and important references.

Let the symbol $\tilde{p}(\cdots \mid \cdots)$ denote the distribution of those random variables listed before the vertical line conditional on those listed after. To obtain a sample from a joint conditional distribution $\tilde{p}(T_1, \dots, T_a \mid T_{a+1}, \dots, T_d)$, the systematic scan Gibbs sampler iterates the following loop.

Sample:

$$(5)\qquad \begin{array}{ll} (1) & T_1^{(m+1)} \text{ from } \tilde{p}\big(T_1 \mid T_2^{(m)}, \dots, T_a^{(m)}, T_{a+1}, \dots, T_d\big). \\ (2) & T_2^{(m+1)} \text{ from } \tilde{p}\big(T_2 \mid T_1^{(m+1)}, T_3^{(m)}, \dots, T_a^{(m)}, T_{a+1}, \dots, T_d\big). \\ & \vdots \\ (\mathrm{a}) & T_a^{(m+1)} \text{ from } \tilde{p}\big(T_a \mid T_1^{(m+1)}, \dots, T_{a-1}^{(m+1)}, T_{a+1}, \dots, T_d\big). \end{array}$$

If the algorithm converges, for a sufficiently large value of m the distribution of $T_1^{(m)}, \dots, T_a^{(m)}$ approximates the equilibrium distribution $\tilde{p}(T_1, \dots, T_a \mid T_{a+1}, \dots, T_d)$ of the Markov chain. Independently replicating this Markov chain l times, and only retaining the last realization each time, produces an independent and identically distributed sample of size l from the approximating distribution. Generally, however, practitioners use more than just the final chain iteration [Geyer (1992) and Gelman and Rubin (1992)].

2.3. *Conditional cumulative distribution function approximations.* Often the one-dimensional marginal distributions of (5) are unavailable. Kolassa and Tanner (1994) suggest instead sampling from approximations to the appropriate conditional cumulative distribution functions. Attention will be primarily focused on two double saddlepoint approximations. The double saddlepoint cumulative distribution function approximation of Skovgaard (1987) generalizes the approximation due to Lugannani and Rice (1980). Suppose a vector $\mathbf{T}$ arises from the regression model specified by (1), (2) and (3), with (1) representing a density with respect to Lebesgue measure, and an approximation is desired for conditional probabilities specified by components of $\mathbf{T}$. Suppose further that $\mathbf{Z}$ is comprised of n repetitions of a fixed set of covariate vectors $\mathbf{W}$, divided by n, and that the approximation is desired to be asymptotic in n. Skovgaard (1987) gives the double saddlepoint approximation as

$$F^1(T_u \mid \mathbf{T}_{-u}) = \Phi\big(\sqrt{n}\,\hat{w}_u\big) + n^{-1/2}\phi\big(\sqrt{n}\,\hat{w}_u\big)\left(\frac{1}{\hat{w}_u} - \frac{1}{\check{z}_u}\right), \tag{6}$$

where $\mathbf{T}_{-j}$ denotes the vector $\mathbf{T}$ with component j deleted, and

$$\check{z}_u = \rho_u^* \hat{\beta}_u, \qquad \rho_u^* = \sqrt{\left|-\ell_0''(\hat{\beta})\right|}\Big/\sqrt{\left|-\ell_{0,-u}''(\tilde{\beta})\right|},$$
$$\hat{w}_u = \operatorname{sgn}(\hat{\beta}_u)\sqrt{2\big[\ell_0(\hat{\beta}) - \ell_0(\tilde{\beta})\big]}.$$

Here ℓ_0 is the log likelihood for the regression model associated with $\mathbf{W}$ and $\mathbf{t}$, and $\hat{\beta}$ and $\tilde{\beta}$ solve

$$\ell_0^j(\hat{\beta}) = 0 \;\forall\, j \quad \text{and} \quad \ell_0^j(\tilde{\beta}) = 0 \;\forall\, j \neq u,\ \tilde{\beta}_u = 0,$$

$\ell_{0,-u}''$ is the $(d-1)\times(d-1)$ submatrix of the matrix of second derivatives of ℓ_0, corresponding to all components of β and $\mathbf{T}$ except component u, and Φ and ϕ are the normal distribution function and density, respectively. Jensen (1992) shows that (6) differs by a term of relative size $O(1/n)$ from

$$F^2(T_u \mid \mathbf{T}_{-u}) = \Phi\big(\sqrt{n}\,\hat{w}_u + \log(\check{z}_u/\hat{w}_u)/(\sqrt{n}\,\hat{w}_u)\big), \tag{7}$$

and so approximations (6) and (7) both have relative error of size $O(n^{-1})$.

When (1) represents a mass function and $\mathcal{Y}$ is a lattice of equally spaced points (which without loss of generality will be taken as $\mathbb{Z}$), Skovgaard (1987) derives a counterpart of (6), in which $\hat{\beta}_u$ is replaced by $2\sinh(\frac{1}{2}\hat{\beta}_u)$ in the definition of $\check{z}_u$, and in which t_u is corrected for continuity when calculating $\hat{\beta}$. That is, if possible values for T_u are 1 unit apart, $\hat{\beta}$ solves $\ell'(\hat{\beta};\tilde{\mathbf{t}}) = 0$ where $\tilde{t}_j = t_j$ if $j \neq u$ and $\tilde{t}_u = t_u - \frac{1}{2}$. The same correction also applies to (7).

Results about the accuracy of the resulting equilibrium distribution will require that approximating steps in the constructed Gibbs sampler be close to those in the sampler using the exact, but unknown and difficult to calculate, conditional distributions, and so the difference in probabilities assigned by approximations (6) and (7) must be considered. The derivatives of (6) and (7) have the form

$$f^j_{T_u \mid \mathbf{T}_{-u}}(\mathbf{t}) = \check{f}_{T_u \mid \mathbf{T}_{-u}}(\mathbf{t})\big(1 + \check{b}^j_{u,n}(\mathbf{t})/n\big) \tag{8}$$
$$= f_{T_u \mid \mathbf{T}_{-u}}(\mathbf{t})\big(1 + b^j_{u,n}(\mathbf{t})/n\big) \tag{9}$$

for $\check{f}_{T_u \mid T_{-u}}(\mathbf{t}) = \sqrt{n}\,\phi(\sqrt{n}\,\hat{w}_u)\rho_u^*$, and $j \in \{1,2\}$. Routledge and Tsao (1997) discuss the approximation $f^1_{T_u \mid \mathbf{T}_{-u}}$ in detail. Kolassa (1998a, b) proves that if the standardized third moments of the distributions in (1) are bounded, there exist constants κ and n_0 such that for $j \in \{1,2\}$,

$$-n \le \check{b}^j_{u,n}(\mathbf{t}) \le \kappa \exp\big(n_0 \hat{w}_u^2/2\big) \qquad \forall\, n \ge n_0 \tag{10}$$

and under additional regularity conditions,

$$\left|\mathbf{b}_{u,n}^{j}(\mathbf{t})\right| \le \kappa \exp\bigl(n_0 \hat{w}_u^2/2\bigr) \qquad \forall\; n \ge n_0. \tag{11}$$

The bound (10) will be used to verify that for sufficiently large n, the approximations (6) and (7) are monotonic, and hence can be inverted. It is also used to define small sets as in Section 2.1, and to demonstrate that their probability content is bounded away from zero, hence to demonstrate that the approximate Gibbs sampling Markov chain converges geometrically. The bound (11) will be used to demonstrate that the resulting equilibrium distribution approximates the desired distribution to high order.

3. The proposed chain. Formally define the Markov chain by setting $\mathbf{t}^0$ equal to the observed vector of sufficient statistics, and for each $u \in \{1, \dots, a\}$ setting

$$T_u^{(m+1)} = F^{\dagger-1}\bigl(V_u^{(m+1)}; T_1^{(m+1)}, \dots, T_{u-1}^{(m+1)}, T_{u+1}^{(m)}, \dots, T_d^{(m)}\bigr), \tag{12}$$

where $V_u^{(m+1)}$ is drawn from a uniform population independently of earlier chain steps, and $F^{\dagger}$ is a conditional cumulative distribution function approximation. We simulate (5), but at each step replace the exact conditional distribution function by an asymptotic approximation. One might use either (6) or (7), or any approximation satisfying (8)–(11), in this role. Approximations to be considered are all monotone for appropriately large n, and so the inverse will be well defined. Call this chain the "approximating chain," and the chain formed from F rather than $F^{\dagger}$ the "exact chain."

In what follows, irreducibility of the proposed chain is first assessed, and then ergodicity is demonstrated.

3.1. *Irreducibility of chains for certain regression models.* This section considers certain regression models, and first determines when the exact chain defined above applied to the canonical sufficient statistics (2) is irreducible. Suppose responses $\mathbf{y}^{\circ}$ are observed, and again inference on parameters $\{1, \dots, a\}$ is desired, conditional on canonical sufficient statistics associated with the nuisance parameters. Let $\mathbf{U}$ be the last $d - a$ columns of $\mathbf{Z}$. Each step in this chain takes values in the set $\mathfrak{T}_n = \{\mathbf{Z}^{\mathsf{T}}\mathbf{y} \mid y_j \in \mathfrak{Y}_j, \mathbf{U}^{\mathsf{T}}\mathbf{y} = \mathbf{U}^{\mathsf{T}}\mathbf{y}^{\circ}\}$; $\mathfrak{T}_n$ depends on n because $\mathbf{Z}$ depends on n. Irreducibility of the approximating chain is added at the end of this section. Continuous cases are the simplest and are considered first.

THEOREM 3.1. *Suppose that the statistics* $\mathbf{T}$ *are defined as in* (2), *where each* $\mathfrak{Y}_j$ *is a connected open subset of* $\mathbb{R}$, *and that each* Y_j *has a positive density with respect to Lebesgue measure on this space. Then a Gibbs sampling scheme generates an irreducible Markov chain.*

PROOF. Take $\mathbf{t}^0$ and $\mathbf{t}^1$ in $\mathfrak{T}_n$. Since $\mathfrak{T}_n$ is open, there exists N an integer such that $\prod_j (t_j^i \pm (t_j^2 - t_j^1)/N) \subset \mathfrak{T}_n$, for $i = 1, 2$. Then let $\mathbf{v}^{0,0} = \mathbf{t}^1$ and for each m and each $j \in \{1, \dots, a\}$, let $\mathbf{v}^{m,j} = \mathbf{v}^{m,j-1} + (t_j^2 - t_j^1)N^{-1}\boldsymbol{\varepsilon}_j$, where $\boldsymbol{\varepsilon}_j$ is the vector of all zeros except with 1 as component j.

Let $\mathbf{v}^{m+1,0} = \mathbf{v}^{m,a}$. For each $m < N$, $\mathbf{v}^{m,0}$ lies on the segment joining $\mathbf{t}^0$ and $\mathbf{t}^1$, and hence by convexity of $\mathfrak{T}_n, \prod_j (v_j^{m,0} \pm (t_j^2 - t_j^1)/N) \subset \mathfrak{T}_n$. Hence the vectors $\mathbf{v}^{m,j}$ all lie in $\mathfrak{T}_n$. Furthermore, $\mathbf{v}^{N,0} = \mathbf{t}^1$. Hence there exists a set of transitions from $\mathbf{t}^0$ to $\mathbf{t}^1$ having a positive transition density. □

Irreducibility may fail for certain simple discrete cases. Consider the following discrete example: $\mathfrak{Y}_j$ is a subset of the integers from 1 to $i - 1$ for each j. Let $\mathbf{Z}$ be the $d \times K$ matrix with $(1, i, i^2, \dots, i^{K-1})$ and $(1, i + 1, (i + 1)^2, \dots, (i + 1)^{K-1})$ as two columns, and the rest arbitrary. Conditioning on the sufficient statistic associated with either of these two columns in effect conditions on all of the $\mathbf{Y}$, since the $\mathbf{Y}$ can be reconstructed from each of these sufficient statistics by themselves. Hence the chain is reducible, and Gibbs sampling will not yield a sample that is a reasonable representation of the conditional distribution of interest, for the reasons discussed in Section 2.1.

Consider a second logistic regression example, presented in Table 1, in which the first and last components of the sufficient statistic are conditioned on. No series of rearrangements of the indicators in $\mathbf{v}$, each keeping the first, last and second or third components of $\mathbf{s}$ fixed, will draw $\mathbf{s}$ closer to $\mathbf{t}$.

One might be tempted to try to extend the argument of Theorem 3.1 to discrete distributions that will hold asymptotically as the density of discrete points of $\mathfrak{T}_n$ increases. However, such applications usually involve positive probability on the boundary points of the sets $\mathfrak{Y}_j$, which may leave boundary points of $\mathfrak{T}_n$ not communicating with other state space points.

Instead, combinatoric arguments examining rearrangements of the counts in $\mathbf{y}$ are necessary. Two lemmas follow, with conditions including the following:

(13) Each $\mathfrak{Y}_j$ has at least two elements and consists of consecutive integers.

TABLE 1
Logistic regression example leading to a reducible Markov chain

Index					***N***	**y**	**v**
1	1	0	0	0	1	1	0
2	1	1	0	0	0	0	0
3	1	0	1	0	0	0	0
4	1	1	1	0	1	0	1
5	1	0	0	1	1	0	0
6	1	1	0	1	0	0	0
7	1	0	1	1	1	0	0
8	1	1	1	1	1	0	0
$\mathbf{s} = \mathbf{Z}^{\mathsf{T}}\mathbf{y}$	1	0	0	0			
$\mathbf{t} = \mathbf{Z}^{\mathsf{T}}\mathbf{v}$	1	1	1	0			

(14) All items in the last column of $\mathbf{Z}$ are identical and nonzero, and other columns of $\mathbf{Z}$ consist of at most two distinct values.

(15) There exists a path through the rows of $\mathbf{Z}$, where two rows are connected if they are identical except for one entry.

The first concerns possible rearrangements of these counts with their total held constant. The second, doing most of the work of this section, considers modification to these rearrangements necessary to keep some of the sufficient statistics constant. These results depend on characteristics of the design matrix $\mathbf{Z}$; specifically, they depend on the existence of paths through the rows of $\mathbf{Z}$. A theorem follows which summarizes these results. This section concludes with a corollary giving a sufficient condition for irreducibility.

LEMMA 3.2. *If conditions* (13)–(15) *hold for the statistics* $\mathbf{T}$ *of* (2), *then for any* k *and* l *such that* $y_l > 0$ *and* $y_k < \max(\mathfrak{Y}_k)$, *there is a series of changes to* $\mathbf{y}$ *that causes only one component of* $\mathbf{t} = \mathbf{Z}^\top \mathbf{y}$ *to change at each stage, that never changes* t_1, *and that will result in* y_k *being increased by one,* y_l *being decreased by one and all others returned to their initial values.*

PROOF. Let $(g_1, \dots, g_i)$ be the indices for the path connecting row $g_1 = k$ to row $g_i = l$ as in condition (15). It will be shown that through a series of movements in which y_{g_j} is increased and $y_{g_{j+1}}$ is decreased, y_k will be increased by one and y_l will be decreased by one and all other counts y_{g_j} will be returned to their initial values. This will be enough to prove the lemma. Inductively, if $i = 2$, then the result is trivial. Suppose the result holds for paths of length no longer than $i - 1$. If $y_{g_{i-1}} = 0$ then set $y_{g_{i-1}} = 1$, decrease y_{g_i} by one and then apply the induction hypothesis to the series $(g_1, \dots, g_{i-1})$. If $y_{g_{i-1}} > 0$, then apply the induction hypothesis to the series $(g_1, \dots, g_{i-1})$, increase $y_{g_{i-1}}$ by one and decrease y_{g_i} by one. The result follows by induction. □

LEMMA 3.3. *For the statistics* $\mathbf{T}$ *of* (2), *assume conditions* (13)–(15) *and*

(16) *None of the last* $d - a$ *components of* $\mathbf{T}$ *are at their extreme values.*

Then for any k *and* l *such that* $y_l > 0$ *and* $y_k < \max(\mathfrak{Y}_k)$, *and such that rows* k *and* l *agree on their first* a *elements, there is a series of changes to* $\mathbf{y}$, *that causes only one component of* $\mathbf{t} = \mathbf{Z}^\top \mathbf{y}$ *to change at each stage, that never changes any of the sufficient statistics corresponding to components where rows* k *and* l *agree, and that will result in* y_k *being increased by one,* y_l *being decreased by one and all others returned to their initial values.*

PROOF. If $a = d$, this is Lemma 3.2. Let $(g_1, \dots, g_i)$ be the indices for the path connecting all cells, as in condition (15). Suppose that the result holds for $a + 1$. Let m be the number of times this path crosses between rows that differ in entry a. The lemma reduces to the induction hypothesis for a if $m = 0$. Suppose it holds for paths with fewer than m such crossings, and suppose that the current path has m such crossings. Choose k^*, $k^\dagger$ and l^*, $l^\dagger$ representing two pairs of rows in the chain between which covariate a changes, in different directions, such that k^* precedes l^*.

If that portion of the path from $k^\dagger$ to l^* contains all zeros, by condition (16) there exists another row c with the same value for covariate a that is positive. By the induction hypothesis on m there exists a sequence of rearrangements after which y_{l^*} is positive. Similarly, if that portion of the path from $k^\dagger$ to l^* contains all counters at their maximal values, there exists a sequence of rearrangements after which $y_{k^\dagger}$ is below its maximal value. Perform these rearrangements, so that without loss of generality one may assume that y_{l^*} is positive and $y_{k^\dagger}$ is below its maximal value.

Now perform the steps of Lemma 3.2 to decrease y_k and increase y_{k^*}. Perform the steps of Lemma 3.2 to decrease $y_{l^\dagger}$ and increase y_l. Now simultaneously increase $y_{l^\dagger}$ and $y_{k^\dagger}$ and decrease y_{l^*} and y_{k^*}. Finally, return the value of y_c, if necessary. By induction the result holds for all m and then for all a. □

THEOREM 3.4. *For the statistics* **T** *of* (2), *under conditions* (13)–(16), *the associated Gibbs sampling Markov chain associated with conditioning on all but the first a entries of* **T** *is irreducible.*

PROOF. If $\mathbf{t} = \mathbf{Z}^\mathsf{T}\mathbf{y}$, $\mathbf{s} = \mathbf{Z}^\mathsf{T}\mathbf{v}$, $\mathbf{s} \neq \mathbf{t}$ and the last $d - a$ components of $\mathbf{s}$ and $\mathbf{t}$ agree, there exist k and l such that $y_k > v_k$, and $y_l < v_l$. It suffices to show that there exists a sequence of rearrangements of $\mathbf{y}$ decreasing y_k and increasing y_l, such that at each stage, only one component of the sufficient statistic vector changes, that none of $y_{a+1}, \dots, y_d$ changes, and such that other components of $\mathbf{y}$ remain unchanged. This follows from Lemma 3.3. □

COROLLARY 3.5. *The result of Theorem* 3.4 *holds if condition* (15) *is replaced by* 3′.

(17) *For each row* **z** *with a nonfixed unit entry*, *say in column j*, *there exists a row* **w** *identical to* **z** *in* **Z**, *except that* **w** *has a zero in column j*, *and these pairs exhaust* **Z**.

Condition (17) implies condition (15) of Theorem 3.4. □

These results apply to the Gibbs sampling from the exact conditional distributions of interest. These results also apply to the approximate Gibbs scheme when the approximation attaches positive conditional probability whenever the exact conditional probability is positive. Relation (10) insures that this holds for sufficiently large n in approximations (6) and (7) whenever the standardized third moments of the distributions in (1) are bounded. Whether n is large enough can be evaluated practically by noting whether any of the univariate conditional distribution function approximations are nonincreasing. Hence the irreducibility results apply to approximating chains more general than those formed from saddlepoint approximations.

3.2. *Convergence of the Markov chain.* Convergence of the Markov chain constructed by Kolassa and Tanner (1994) will be demonstrated by showing that certain sets are small and by using convergence criteria given by Nummelin (1984) to demonstrate the existence of an equilibrium distribution. Let μ_n be a measure of $\mathfrak{T}^*$ with respect to which the probability distributions (3) are absolutely continuous; μ_n will typically be multivariate Lebesgue measure or counting measure on a lattice with spacings $1/n$. Small sets for this chain may now be defined with reference to derivatives of $\tilde{F}$. For each j, and each random vector $\mathbf{V}$, let $\mathbf{V}^j = \mathrm{E}[\mathbf{V} \mid V_{j+1}, \dots, V_d]$. Let $g_n(\mathbf{t}) = n\mathbf{t}^T\hat{\beta}(\mathbf{t}) - K(\hat{\beta}(\mathbf{t})) - {}^T\beta^\ddagger(\mathbf{t}) + K(\beta^\ddagger(\mathbf{t}))]$, where $\hat{\beta}$ is the maximum likelihood estimator for the regression model, and $\beta^\ddagger$ is the maximum likelihood estimator for the reduced model with the first a columns of $\mathbf{Z}$ removed. Let $\mathscr{T}(\delta) = \{\mathbf{t} \mid g_n(\mathbf{t}) \leq \delta\}$. $\mathscr{T}(\delta)$ and g_n satisfy the requirements of Section 2.1.

LEMMA 3.6. *Construct a Markov chain according to* (5), *where the transitions come from an approximating distribution function with derivative of form* (8) *and with the function* $\check{b}_{u,n}^j$ *satisfying* (10). *Then for any* $\delta > 0$ *there exists* n_0 *such that if* $n > n_0$ *then* $\mathscr{T}(\delta)$ *is a small set for this chain, with a corresponding* α *independent of* n.

PROOF. Demonstrating the above result requires the construction of the measure ν_n. In the continuous case, the transition kernel for this chain is given by

$$\tilde{P}_n(\mathbf{t}, \mathbf{v}) = \left[\prod_{u=1}^{a} \phi\left(\sqrt{n}\,\hat{w}_u\right)\left(\hat{w}_u/\check{z}_u\right)\left(d\hat{w}_u/dt_u\right)\right](1 + b_n(\mathbf{t}, \mathbf{v})/n),$$

where $b_n(\mathbf{t}, \mathbf{v}) = n[\prod_{u=1}^{a}(1 + b_u(v_u; v_1, \dots, v_{u-1}, t_{u+1}, \dots, t_d)/n) - 1]$. Let $\mathbf{m} = \mathbf{T}^a$, the expectation of $\mathbf{T}$ conditional on all but the first a components. Choose n_0 such that $n \geq n_0$ implies $b_n(\mathbf{m}, \mathbf{m}) > -n_0$. Since $(\hat{w}_u/\check{z}_u)(d\hat{w}_u/dt_u)$ and $b(\mathbf{t}, \mathbf{v})$ are jointly continuous, there exist $\delta > 0$ and a matrix Σ such that for $\mathscr{T}$ as above, then $\inf\{\tilde{P}_n(\mathbf{t}, \mathbf{v})/\exp((n/2)(\mathbf{t} - \mathbf{u})^T\Sigma(\mathbf{t} - \mathbf{u})) \mid n > n_0,\ (\mathbf{t}, \mathbf{v}) \in \mathscr{T}(\delta/\sqrt{n}) \times \mathscr{T}(\delta/\sqrt{n})\} > 0$. For a Borel set $\mathscr{H} \subset \mathfrak{T}_n$ let

$$\nu^*(\mathscr{H}) = \int_{\mathscr{H} \cap \mathscr{T}(\delta/\sqrt{n})} \sqrt{n}\, \exp\left(-\frac{n}{2}(\mathbf{t} - \mathbf{u})^T\Sigma(\mathbf{t} - \mathbf{u})\right)\mu_n(d\mathbf{t})$$

and

$$\nu_n(\mathscr{H}) = \frac{\nu^*\left(\mathscr{H} \cap \mathscr{T}\left(\delta/\sqrt{n}\right)\right)}{\nu^*\left(\mathscr{T}\left(\delta/\sqrt{n}\right)\right)}. \qquad \square$$

We now construct a function g as in Lemma 2.3 as a preliminary step in demonstrating convergence of the Markov chain.

LEMMA 3.7. *Under the conditions of Lemma* 3.6, *there exists* $\delta > 0$ *such that for sufficiently large* n *the pair* g_n *and* $\mathscr{T}(\delta)$ *satisfy* (4), *with* $\omega_n > 1$, $\gamma_n > 0$, *and* Γ_n *all independent of* n.

PROOF. Let $C = \sup_{n \geq n_0, 1 \leq j \leq a, t \in \mathfrak{T}_n} (\mathrm{E}[\hat{w}_j(\mathbf{T}^j) \mid \mathbf{T}_{j+1} = t_{j+1}, \dots, T_d = t_d] - 1/n)/n$. Under (11), for n_0 suitably large, C is bounded. Then

$$\begin{aligned} \mathrm{E}[g_n(\mathbf{T}) \mid \mathbf{V}] &\leq \frac{n}{2} \sum_j \mathrm{E}\Big[\mathrm{E}\Big[\hat{w}_j(\mathbf{T}^j)^2 \mid T_{j+1}, \dots, T_d\Big] \mid \mathbf{V}\Big] \\ &\leq \frac{1}{2} \sum_j \mathrm{E}\left[1 + \frac{C}{n} \mid \mathbf{V}\right] \leq \frac{a}{2} + \frac{aC}{2n}. \end{aligned}$$

Choose δ large enough so that n sufficiently large, $\inf_{v \notin \mathscr{T}(\delta/\sqrt{n})} \geq a/2 + 1$. □

THEOREM 3.8. *The Markov chain defined in Lemma* 2.3 *is geometrically ergodic*, *for sufficiently large* n.

The convergence properties follow by applying the results, Lemmas 2.3 and 3.7 to the small set $\mathscr{T}$ and the positive function g defined above.

While aperiodicity is difficult to prove in general, it is quite easy to demonstrate in practice. One need only exhibit a chain state $\mathbf{t}$ for which $\tilde{P}_n(\mathbf{t}, \mathbf{t}) > 0$. Lemma 3.6 used the fact that for sufficiently large n this holds at $\mathbf{m}$. In all examples we have investigated $\tilde{P}_n(\mathbf{t}, \mathbf{t}) > 0$ for the observed vector of sufficient statistics, but in general one could check all of the sampled vectors until such a point is found.

Tierney (1994) discusses strategies for demonstrating uniform and geometric ergodicity. Schervish and Carlin (1992) discuss strategies for demonstrating that a Markov chain is geometrically ergodic and provide a simple counterexample demonstrating that uniform ergodicity is not to be expected from the Gibbs sampler.

Since the unidimensional sampling steps are not exactly the conditional distributions arising from the resulting equilibrium distribution, this equilibrium distribution may depend on the order in which sampling is done. The following section will show, however, that all of these alternate equilibrium distribution will well approximate the desired multivariate conditional distribution.

3.3. *Accuracy of the equilibrium distribution.* Let μ_n be as in Section 3.2, let $\mathscr{H} = L^1(\mu_n)$, and let $\mathscr{H}_0 = \{g \in \mathscr{H} \mid \int g \, d\mu_n = 0\}$. For a transition density P_n formed from Gibbs sampling in (5), let $S_n^m : \mathscr{H} \to \mathscr{H}$ be the operator mapping a measure g to the measure

$$(S_n^m g)(A) = \int_{\mathscr{T}_n} \{P_n^{(m)}(\mathbf{v}, A) - \pi_n(A)\} g(\mathbf{v}) \mu_n(d\mathbf{v}). \tag{18}$$

Heuristically this operator maps an unconditional distribution on the state space for the initial value of chain iterations to the distribution after m iterations of the chain. Similarly, let $\tilde{P}_n^{(m)}$ be the transition density associated with Gibbs sampling as in (5), with observations drawn from (6) rather than the exact full conditional distribution, and let $\tilde{S}_n^m$ be as in (18) with $\tilde{P}_n^{(m)}$

replacing $P^{(m)}$. Form $\check{P}_n(\mathbf{t}, \cdot)$ by taking the pointwise minimum for the densities representing $\tilde{P}_n(\mathbf{t}, \cdot)$ and $P_n(\mathbf{t}, \cdot)$, multiplied by a rescaling factor to force it to integrate to unity. Let $\check{S}_n$ be the associated operator. Straightforward calculations demonstrate that S_n^m and $\tilde{S}_n^m$ result from iterations of S_n^1 and $\tilde{S}_n^1$. Let $\tilde{D}_n$ be the operator $(a\tilde{D}_n g)(A) = \int_{\mathfrak{T}_n}\{\tilde{P}_n(\mathbf{v}, A) - \check{P}_n(\mathbf{v}, A)\}g(d\mathbf{v})$, and let D_n be the corresponding operator with P_n in place of $\tilde{P}_n$. Since $S_n\pi_n = \tilde{S}_n\tilde{\pi}_n = 0$, then

$$(I - \check{S}_n)(\tilde{\pi}_n - \pi_n) = \tilde{D}_n\tilde{\pi}_n + D_n\pi_n, \tag{19}$$

where I is the identity operator.

When $\|S_n\|_{\mathscr{H}_0 \to \mathscr{H}_0} < 1$, and $\|\cdot\|_{\mathscr{H}_0 \to \mathscr{H}_0}$ is the norm on the space of operators from $\mathscr{H}_0$ into itself, then $I - S_n$ has a continuous inverse, and $\pi_n - \tilde{\pi}_n$ may be isolated by multiplying by this inverse. The condition $\|S_n\|_{\mathscr{H}_0 \to \mathscr{H}_0} < 1$ is implied by $\inf\{P_n(\mathbf{t}, \mathbf{v}) \mid t \in \mathfrak{T}_n\} > 0$ for all $\mathbf{t} \in \mathfrak{T}_n$, as demonstrated by Roberts and Polson (1994). The following theorem addresses cases in which this minorization condition does not hold.

THEOREM 3.9. *If the Markov chain with transition kernel $\tilde{P}_n$ is irreducible and if the approximating chain is constructed as in Section* 3, *with the asymptotic approximation used in* (12) *satisfying* (8)–(11), *then the equilibrium distribution for the approximating chain satisfies $\|\pi_n - \tilde{\pi}_n\|_{\mathrm{TV}} = O(1/n)$. Furthermore, these conditions are met by the asymptotic approximations* (6) *and* (7).

PROOF. Conditions (8)–(11) imply that there exists n_0 such that $\tilde{P}_n(\mathbf{v}, \mathbf{t}) - P_n(\mathbf{v}, \mathbf{t}) \mid$ is bounded by a constant times $\sqrt{n/(n - n_0)}\, P_{n-n_0}(\mathbf{v}, \mathbf{t})$ if $n > n_0$. The uniformity of the double saddlepoint density approximation implies that the density of $\pi_n(\mathbf{t})$ is bounded above and below by constants times $\sqrt{n/(n - n_0)}\, \pi_{n-n_0}(\mathbf{t})$, then $\tilde{D}_n\tilde{\pi}_n$ is bounded by a constant times $\tilde{\pi}_n$.

Choose any integer $h > 0$. Multiply (19) by $\sum_{j=0}^h \check{S}_n^j$, to obtain $(I - \check{S}_h^{h+1})(\tilde{\pi}_n - \pi_n) = \sum_{j=0}^h \check{S}_n^j \times [\tilde{D}_n\tilde{\pi}_n + D_n\pi_n]$, and $\tilde{\pi}_n - \pi_n = \check{S}_n^{h+1}(\tilde{\pi}_n - \pi_n) + \sum_{j=0}^h \check{S}_n^j[\tilde{D}_n\tilde{\pi}_n + D_n\pi_n]$. Let σ_t be the measure assigning probability one to $\mathbf{t}$. Hence $\|\tilde{\pi}_n - \pi_n\|_{\mathrm{TV}} \le \int \|\check{S}_n^{h+1}\sigma_t\|_{\mathrm{TV}}(\tilde{\pi}_n + \pi_n)(d\mathbf{t}) + \sum_{j=0}^h \check{S}_n^j[\tilde{D}_n\tilde{\pi}_n + D_n\pi_n]$. Since $\check{S}_n < \sqrt{\omega_1}\, S_n$ and $\check{S}_n < \sqrt{\omega_1}\, \tilde{S}_n$ if $n > B_0/(\sqrt{\omega_1} - 1)$, then for such n, $\|\tilde{\pi}_n - \pi_n\|_{\mathrm{TV}} \le \omega_1^{-h/2} \int [\tilde{\psi}_n(\mathbf{t})\tilde{\pi}_n(d\mathbf{t}) + \psi_n(\mathbf{t})\pi_n(d\mathbf{t})] + \sum_{j=0}^h \|\check{S}_n^j[\tilde{D}_n\tilde{\pi}_n + D_n\pi_n]\|_{\mathrm{TV}}$, and the equilibrium density $\check{\pi}_n$ for $\check{P}_n$ satisfies $\check{\pi}_n \le \nu_n G_{s,\nu}^{(\omega_1)}$. Choose ω as in Lemma 3.7 and choose $\tilde{\psi}_n$ as in Lemma 2.3, corresponding to the chain with kernel P_n. Note that

$$\tilde{D}_n \le \frac{\sqrt{n - n_0}}{\sqrt{n}} \frac{\kappa}{(n - n_0 + \kappa)} \tilde{S}_{n-n_0}$$

and

$$\tilde{S}_{n-n_0} \le \frac{\sqrt{n - n_0}}{\sqrt{n}} \frac{(n + \kappa)}{(n - n_0 + \kappa)} \tilde{S}_n,$$

with the same relationships holding for $D_.$ and $D_.$ in place of $\tilde{D}_.$ and $\tilde{S}_.$. Hence for sufficiently large n, $\tilde{D}_n \le n^{-1}\sqrt{\omega_1}\,\tilde{S}_n$ and $D_n \le n^{-1}\sqrt{\omega_1}\,S_n$. Then $\|\tilde{\pi}_n - \pi_n\|_{\mathrm{TV}} \le \omega_1^{-h/2}\int[\tilde{\psi}_n(\mathbf{t})\tilde{\pi}_n(d\mathbf{t}) + \psi_n(\mathbf{t})\pi_n(d\mathbf{t})] + n^{-1}\sum_{j=0}^{h} \omega_1^{j/2}\|\tilde{S}_n^j\tilde{\pi}_n + S_n^j\pi_n\|_{\mathrm{TV}}$. Let $h \to \infty$ to find

$$\begin{aligned}\|\tilde{\pi}_n - \pi_n\|_{\mathrm{TV}} &\le n^{-1} \sum_{j=0}^{\infty} \omega_1^{j/2}\Big(\|\tilde{S}_n^j\tilde{\pi}_n\|_{\mathrm{TV}} + \|S_n^j\pi_n\|_{\mathrm{TV}}\Big)\\ &\le n^{-1}\int\Big[\tilde{\psi}_n(\mathbf{t})\tilde{\pi}_n(d\mathbf{t}) + \psi_n(\mathbf{t})\pi_n(d\mathbf{t})\Big];\end{aligned}$$

the integral above is bounded by Lemma 2.3.

The suitability of (6) and (7) was discussed in Section 2.2. □

Acknowledgments. The author gratefully acknowledges suggestions from Martin Tanner, Yodit Seifu, Fengchung Peng, the organizers and participants of the Interface session at which this work was presented and the organizers and participants at the AMS-IMS-SIAM conference on Markov chain Monte Carlo methods.

REFERENCES

DIACONIS, P. and STURMFELS, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363–397.

FORSTER, J. J., MCDONALD, J. W. and SMITH, P. W. F. (1996). Monte Carlo exact conditional tests for log-linear and logistic models. *J. Roy. Statist. Soc. Ser. B* **58** 445–453.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GEYER, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–483.

HIRJI, K. F., MEHTA, C. R. and PATEL, N. R. (1987). Computing distributions for exact logistic regression. *J. Amer. Statist. Assoc.* **82** 1110–1117.

JENSEN, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika* **79** 693–703.

KOLASSA, J. E. (1998a). Uniformity of double saddlepoint conditional probability approximations. *J. Multivariate Anal.* **64** 66–85.

KOLASSA, J. E. (1998b). Bounding the difference between two saddlepoint distribution function approximations. Technical Report 98-04, Dept. Biostatistics, Univ. Rochester.

KOLASSA, J. E. and TANNER, M. A. (1994). Approximate conditional inference in exponential families via the Gibbs sampler. *J. Amer. Statist. Assoc.* **89** 697–702.

LUGANNANI, R. and RICE, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.* **12** 475–490.

MEHTA, C. R., PATEL, N. R. and SENCHAUDHURI, P. (1993). Monte Carlo methods for conditional logistic regression. In *Computer Science Statistics. Proceedings of the 25th Symposium on the Interface* (Michael E. Tarter and Michael D. Lock, eds.) **25** 385–391. Interface Foundation of North America, Fairfax Station, VA.

NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Univ. Press.

ROBERTS, G. O. and POLSON, N. G. (1994). On the geometric convergence of the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **56** 377–384.

ROBERTS, G. O. and SMITH, A. F. M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms. *Stochastic Process. Appl.* **49** 207–216.

ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90** 558–566.

ROUTLEDGE, R. and TSAO, M. (1997). On the relationship between two asymptotic expansions for the distribution of sample mean and its applications. *Ann. Statist.* **25** 2200–2209.

SCHERVISH, M. J. and CARLIN, B. P. (1992). On the convergence of successive substitution sampling. *J. Comput. Graph. Statist.* **1** 111–127.

SKOVGAARD, I. M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.* **24** 875–887.

TANNER, M. A. (1996). *Tools for Statistical Inference*. Springer, Berlin.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762.

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF ROCHESTER MEDICAL CENTER
SCHOOL OF MEDICINE AND DENTISTRY
601 ELMWOOD AVE., BOX 630
ROCHESTER, NEW YORK 14642
E-MAIL: kolassa@bio1.bst.rochester.edu